Personnel Assessment and Decisions

Volume 7 | Issue 2 Article 1

2021

Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action

Nancy T. Tippins
The Nancy T. Tippins Group, LLC

Frederick L. Oswald
Rice University

S. Morton McPhail
Independent Researcher

Recommended Citation

Tippins, Nancy T.; Oswald, Frederick L.; and McPhail, S. Morton (2021) "Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action," *Personnel Assessment and Decisions*: Number 7 : Iss. 2 , Article 1.
DOI: https://doi.org/10.25035/pad.2021.02.001
Available at: https://scholarworks.bgsu.edu/pad/vol7/iss2/1

Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action

Nancy T. Tippins[1], Frederick L. Oswald[2], and S. Morton McPhail[3]

1. The Nancy T. Tippins Group, LLC
2. Rice University, Department of Psychological Sciences
3. Independent Researcher

ABSTRACT

Organizations are increasingly turning toward personnel selection tools that rely on artificial intelligence (AI) technologies and machine learning algorithms that, together, intend to predict the future success of employees better than traditional tools. These new forms of assessment include online games, video-based interviews, and big data pulled from many sources, including test responses, test-taking behavior, applications, resumes, and social media. Speedy processing, lower costs, convenient access, and applicant engagement are often and rightfully cited as the practical advantages for using these selection tools. At the same time, however, these tools raise serious concerns about their effectiveness in terms of their conceptual relevance to the job, their basis in a job analysis to ensure job relevancy, their measurement characteristics (reliability and stability), their validity in predicting employee-relevant outcomes, their evidence and normative information being updated appropriately, and the associated ethical concerns around what information is being represented to employers and told to job candidates. This paper explores these concerns, concluding with an urgent call to industrial and organizational psychologists to extend existing professional standards for employment testing to these new AI and machine learning based forms of testing, including standards and requirements for their documentation.

KEYWORDS

personnel selection, artificial intelligence, big data, games, video-based interviews, test standards

To say today that technology is deeply embedded in tests and assessments is clichéd. For at least 30 years, technology has been widely used to facilitate tests and assessments in employment settings in one form or another, and the benefits and liabilities of technology-based employment testing have been well-documented (e.g., Tippins & Adler, 2011). In a nutshell, the increased speed and cost reductions provided by technology are particularly appealing to employers who seek to evaluate a large pool of job candidates and select from it in a timely and effective manner. In turn, candidates themselves have increasingly come to expect technologically current, convenient, and engaging selection processes. Despite the manifest and potential benefits of new technologies, many challenges remain, such as identifying qualified candidates reliably and preventing cheating and other forms of malfeasance that threaten the integrity of assessment results.

In this paper, we want first to draw the reader's attention to the challenges associated with new forms of testing, including those that use artificial intelligence, which are growing in popularity. Second, we want to make a plea to the community of industrial and organizational (I-O) psychologists to extend the existing standards to these new technologies (i.e., the *Principles for the Validation and Use of Personnel Selection Procedures* [*Principles*] and the *Standards for Educational and Psychological Testing* [*Standards*]). We recognize that virtually every person trained in I-O psychology is aware of the foundational importance of

Corresponding author:
Nancy T. Tippins
Email: nancy@tippinsgroup.com

Authors contributed equally to the manuscript. We welcome your feedback.

reliability, validity, and fairness; however, many do not understand how to evaluate these fundamental characteristics when these new methods are used. Compounding this issue, many of our employers and clients look to us for guidance on how to best use these emerging tools and evaluate them appropriately.

New Forms of Assessment

Employment testing often connotes the use of a structured instrument to collect responses from a test taker that, when scored, would indicate his/her standing on the construct being measured. Almost everyone is familiar with common forms of testing to assess competence in wide-ranging subject areas using multiple-choice, fill-in-the-blank, essay, and matching formats. In the U.S., the *Uniform Guidelines on Employee Selection Procedures* (*Uniform Guidelines*; EEOC, CSC, DoL, DoJ, 1978) broadened the meaning of the word *test* in employment settings to include any selection procedure used as the basis for an employment decision:

> These guidelines apply to tests and other selection procedures which are used as a basis for any employment decision. Employment decisions include but are not limited to hiring, promotion, demotion, membership (for example, in a labor organization), referral, retention, and licensing and certification, to the extent that licensing and certification may be covered by Federal equal employment opportunity law. Other selection decisions, such as selection for training or transfer, may also be considered employment decisions if they lead to any of the decisions listed above. (*Uniform Guidelines*, 1978, 2B)

The *Uniform Guidelines Questions and Answers* (*Uniform Guidelines*, Q&A, 1979) further expanded the concept of *test* to include "job requirements (e.g., physical, education, experience) and evaluation of applicants on the basis of application forms, interviews, performance tests, paper and pencil tests, performance in training programs or probationary periods, and any other procedures used to make an employment decision" (Question 6). In line with this broadened concept of a "test," we use the words *test* and *assessment* to refer to any form of data that contributes to a selection procedure, as described in detail below.

Historically, computerized testing was conducted in person and was used as little more than a way to save paper and automatically score traditional forms of testing. In the 1980s, numerous tests using multiple choice item formats were converted to computer administration and scoring. Initially, the equivalence of paper-and-pencil and computerized versions of the same test was a primary concern, because both formats were being used in the same selection settings. With time, the boundary conditions affecting test administration (e.g., test content, time constraints, answer entry vs. answer selection) were identified, and the extent of the equivalence of the two formats became better understood. Later, internet accessibility also promoted the increased use of unproctored Internet testing (UIT), which raised new concerns about test taker identity and cheating. More recently, new testing technologies have posed additional challenges for test users related to the equivalency of scores on tests taken on different devices (e.g., mobile and desktop formats) and test taker distraction. Many scientist–practitioners in the field of employment testing have lamented the focus on using technology to manage old ways of testing more effectively if it comes at the expense of using technology to develop new and better ways of assessing job-relevant knowledge, skills, abilities, and other characteristics (KSAOs) (Tippins, 2009).

Yet, in fact, in recent years, technology in the employment testing arena has proliferated and diversified in ways that have resulted in radically different methods for evaluating candidates' qualifications. We recommend a three-part framework for understanding these new technologically enhanced forms of assessment based on different *technologies*, types of *data*, and *algorithms*. Examples of different *technologies* include online games, video interviews, and social media. Technologies are independent from the constructs being measured and therefore should not be confused with them (Arthur & Villado, 2008; Campbell & Fiske, 1959). Most formats can measure a wide range of constructs, and conversely, many constructs can be evaluated with a wide range of technologies. Therefore, a new technology cannot be said to be universally valid. Instead, evidence needs to be marshaled concerning whether job-relevant constructs are being measured, which in turn informs other evidence regarding validity and fairness derived from the test scores and the inferences made from them. The "modular approach" to selection recommends examining these different technologies and sources of information systematically with respect to the constructs being measured (Lievens & Sackett, 2017).

The *data* collected using new technologies may be usefully distinguished along a continuum (Oswald, 2020). Anchoring one end of the continuum are more traditional *intentional* responses to a prompt such as a test item or interview question. On the other end lie more *incidental* data that are less intentional or controllable by the respondents, requiring little or no effort (and in some cases little control) on the applicants' part (e.g., social media posts, facial movements or voice characteristics in a video interview). When these new, less obtrusive technologies are used to collect information electronically, they tend to be more incidental in nature, and the amount of data gathered can be massive. For instance, data from game-based testing might include all mouse clicks, thus capturing decisions, locations, and responses for each scenario; the number of mouse clicks used to achieve a given outcome; and event-based and between-click response times. Video interviews may produce data pertaining to voice quality and facial features that are collected contnuously during the interview, as well as word usage, themes, length of words, use of personal pronouns, and so on. Other "big data" approaches to selection may incorporate a wide range of data pulled from many different sources (e.g., application forms, emails, social media; see Guzzo et al., 2015, for further discussion).

In addition to technologies and data, the third category in the framework is *algorithms*. Artificial intelligence (AI) is a broad term referring to computer-based procedures that mimic the decisions, processes, or outcomes of humans so closely as to appear intelligent. In this context, machine learning (ML) is a subset of AI, referring to the algorithms (mathematical and statistical procedures) underlying these procedures, and deep learning (DL) is a subset of ML, referring to neural-network-based ML algorithms. In this paper, we have used the term *algorithms* somewhat loosely to refer to computational procedures using iterative processes that converge on a "best" set of models and parameter estimates for effective clustering or prediction in new samples of data.

Data going into an ML algorithm can come from a wide variety of information sources, such as text, images, social media, voice quality, and facial features (Oswald et al., 2020). In the prediction context, the "learning" part of machine learning happens when the algorithms are first exposed to large amounts of data—also called the *training set*—to establish predictions of a criterion of interest (e.g., supervisory ratings of performance, judgments about who is a "good" employee, productivity or sales scores, interim outcomes such as referral for further consideration). The resulting "trained" models are only viewed as effective if they make sufficiently accurate predictions on an independent sample—called the *test set*—comprising data that were not used to train the algorithm and develop the model. Test sets are entirely new, independent data sets obtained from a hold-out sample within the existing data set (e.g., a random 20% selected), from the folds within a *k*-fold cross-validation procedure, or from a newly collected data set.

There are hundreds of ML algorithms, and the number continues to grow (see the list of algorithms in the R package caret [Kuhn, 2008] at https://topepo.github.io/caret/available-models.html). In general, ML algorithms fall into two broad types. *Supervised learning* algorithms involve prediction of some criterion, such as when applicant data are used to predict job performance (i.e., the criteria "supervise" how predictors are used). *Unsupervised learning* algorithms involve grouping people or cases into clusters, such as technically versus interpersonally effective performers, or applicants whose characteristics are like or unlike those of high performing incumbents. These methods do not involve the use of criteria, making the grouping process "unsupervised."

Regardless of the type of ML approach used, both supervised and unsupervised algorithms often operate on big datasets, mining large numbers of variables, sometimes from large numbers of people, in an attempt to make accurate predictions or clustering assignments. For most ML algorithms, the number of variables can even exceed the number of cases. In addition, the data may be messy, have missing values, and be used in their raw form. By contrast, traditional analyses, such as linear regression and ANOVA, only function effectively when the number of variables is far fewer than the number of cases. To reduce the number of variables, item composites may be created or some variables may even be excluded. Note that different ML algorithms often make highly similar predictions and end up with similar overall levels of accuracy (Domingos, 2012). For applications of ML in personnel selection settings, we claim that, generally, the effectiveness of ML prediction or clustering will more likely be driven by availability of high-quality data than by which ML algorithm is chosen. Whether the advantages of a large number of predictors offset the disadvantages of "messy" data must be determined in each case. Thus, the expectation of a significant increase in organizational benefit from machine learning used to predict employee performance ratings made by supervisors is not always well-founded. Accurate and well justified predictions depend on good measurement processes and good data as well.

Throughout this paper, we have used the descriptor *technologically enhanced* to refer to the broad sweep of possible assessments involving technologies, data, and algorithms that are associated with new approaches to selection that might involve such data as facial recognition; virtual reality; gamification; massive amounts of data from resumes, applications, social media; and the like. We have also attempted to clarify some of the other terms used, such as *artificial intelligence*, *machine learning*, and *deep learning*. We have risked sounding pedantic in our desire for greater clarity when we raise questions about how relatively recent developments in technologically enhanced selection are to be evaluated, how they may fit into existing and evolving legal frameworks, and what role industrial and organizational (I-O) psychologists can and should play in addressing these concerns.

In addition to a variety of additional concerns discussed below, this paper suggests that many, if not most, AI-based selection tests are deficient in terms of transparency and documentation. There is little to no supplemental information available for many technologically enhanced assessments to indicate what processes were used to develop the tools and how (or if) the assessments provide improvements over traditional measures (e.g., more faithfully reflecting the constructs of interest, having higher reliability, better validity, and meaningful incremental validity). Developers and vendors also do not tend to provide sufficient documentation regarding the typical steps taken to clean data, develop algorithms, evaluate results, and so on, all of which contribute to demosnstrating alignment with ethical, legal, and professional standards. Raghavan et al. (2020) studied the claims made by 18 AI assessment vendors about their algorithmic assessments used for employee selection and concluded, "Transparency is crucial to further our understanding of these systems. While there are some exceptions, vendors in general are not particularly forthcoming about their practices. Additional transparency is necessary to craft effective policy and enable meaningful oversight" (p. 17). Noting the financial incentives for companies to develop AI-based tools, Narayanan (2019), a computer scientist, found that AI was no better than more transparent linear regression models for predicting social outcomes. He concluded, "AI excels at some tasks, but can't predict social outcomes. We must resist the enormous commercial interests that aim to obfuscate this fact. In most cases, manual scoring rules are just as accurate, far more transparent, and worth considering." These statements seem to reinforce the perspective that in terms of reliability, validity, and fairness, it is often not clear whether or how AI assessments are offering added value beyond the more traditional forms of measures, data, and algorithms.

Advantages

These new, technology enhanced assessments have many advantages that are appealing to organizations, although of course, not all benefits of technology apply wholesale to all of types of assessments. Most are very fast and efficient, and they can be easily deployed in remote locations and scored immediately. When the economy is booming and the competition for talent is fierce, the value of speed and efficiency to the employer cannot be overemphasized. One employer considering using a technology enhanced resume screener reported receiving 500,000 resumes annually for its sales jobs. Even if the employer were to spend one minute per resume, it would take over 1,000 staff days just to review all resumes. Obviously, such an approach would be infeasible and ineffective, and smart compromises need to be made.

In addition to speed and efficiency, many technologically enhanced assessments have other advantages. For example, video games may attract and engage candidates for certain jobs and organizations, which may be especially useful in a tight labor market where many employers are competing for talent. Other types of assessments may have the benefit of being unobtrusive by evaluating job candidates on the basis of readily available information; however, candidates may not be aware of all the information being obtained and used (e.g., information on social media), raising both ethical and legal concerns (discussed further below). To the extent such data mining is permissible, it can help organizations locate highly qualified "passive" job candidates who have not expressed any desire to change jobs and are not looking.

Another advantage often expressed by vendors of technology enhanced assessments is a lack of or reduction in adverse impact and/or increases in criterion-related validity, as compared with traditional testing. However, the empirical evidence behind such claims is often unavailable, making relevant comparisons impossible. Absent such evidence, we cannot rest assured that the validity-adverse impact dilemma has been improved, let alone solved (Pyburn et al., 2008).

A related advantage frequently claimed is the reduction or elimination of bias and the expansion of the candidate pool. For example, in video-based interviews, the same questions are asked of all candidates, with no cues as to what is a good response (Weed, 2020). A paper by the cofounder and CEO of an AI-based testing firm (Polli, 2019) suggested two practical advantages for using AI in hiring: (a) Both human bias and biased selection tools contribute to unfair hiring, and artificial intelligence can eliminate this human bias[1]. (b) Large pools of applicants are ignored because traditional approaches cannot evaluate large numbers of candidates, whereas AI-based processes can. We certainly agree with these two points as *possibilities*; yet, again, the available evidence supporting such propositions is scarce to nonexistent. Next-generation AI talent management tools and systems should provide all stakeholders (e.g., HR professionals, I-O psychologists from both the researcher and practice domains, lawyers, ethicists, and, importantly, job applicants) with clearer evidence supporting such assertions.

Disadvantages

The rapid increase in innovative technological approaches for assessing job candidates comes with a similar increase in significant concerns about them and the scarcity of appropriate evidence. The news media have raised legitimate fairness and privacy concerns about many of these approaches. For example, the *Washington Post* reviewed the pros and cons of AI applications of face scanning in hiring systems. Although rightfully acknowledging that human judgment also is rife with inconsistencies, biases, and errors, the article also cites fundamental problems with AI and face scanning[2]:

[1] Of course, this statement depends on the definition of bias (e.g., biased data, biased models, biased outcomes). To the extent that the data used in the model are biased, there is no guarantee that the use of AI tools will reduce or eliminate bias.

[2] It merits noting that one vendor has recently eliminated the evaluation of facial characteristics from their interview models because it deemed that its incremental prediction was insufficient.

> But some AI researchers argue the system is digital snake oil—an unfounded blend of superficial measurements and arbitrary number-crunching that is not rooted in scientific fact. Analyzing a human being like this, they argue, could end up penalizing nonnative speakers, visibly nervous interviewees or anyone else who doesn't fit the model for look and speech.
>
> The system, they argue, will assume a critical role in helping decide a person's career. But they doubt it even knows what it's looking for: Just what does the perfect employee look and sound like, anyway? (Harwell, October 22, 2019)

Adding to this concern about facial-recognition systems, a *New York Times* article (Singer & Metz, 2019) summarized a study by the National Institute of Standards and Technology (NIST) that tested 189 facial-recognition algorithms from 99 developers, concluding that African American and Asian faces are misidentified 10 to 100 times more frequently than Caucasian faces. Although the purpose of the facial recognition systems examined by NIST was different than that of pre-employment selection, the article raises important concerns about such techniques producing differential results and outcomes with respect to race and ethnicity. Another study called Gender Shades, reported in the *Wall Street Journal* (2020), found that three facial-recognition systems were much less likely to correctly identify the faces of darker skinned women compared to light-skinned men, with error rates of 1% vs. 35%, respectively.

In reporting on a complaint to the Federal Trade Commission (FTC) by the Electronic Privacy Information Center (EPIC), *The Washington Post* (Harwell, November 6, 2019) cited several problems with artificial intelligence systems in the hiring process: (a) Candidates are not provided information about their scores, and (b) they are unaware that their personal data are being used for evaluation purposes.

Others have raised questions about the extent to which machine learning techniques (especially, supervised methods) themselves may incorporate bias into the resulting prediction (Illingworth, 2015). Such concerns raise numerous questions, such as whether and when the underlying data include embedded bias, how it should be addressed, and the importance of being able to explicate the complex algorithmic results in terms that demonstrate job relatedness (see discussion below).

These recent complaints, as well as many of the claims made by test publishers about technology enhanced assessments, should be carefully investigated, verified, and vigorously debated. Most I-O psychologists would be reluctant to accept at face value assertions that the reliability, validity, and adverse impact of technologically enhanced selection tools are as good as or better than their traditional counterparts without examining empirical evidence. Yet, as we have already noted, such evidence is often lacking.

To be clear, it is not our intent to evaluate the legitimacy of the purported benefits and potential problems of any specific technologically enhanced selection tool. Rather, our purpose is broader—to point out that from many quarters (e.g., scientific, public, media, and legal), there is a growing desire for such assessments to be evaluated against testing and assessment standards that, across many decades, have been relevant to any other form of testing. The potential for reduced or no adverse impact in an AI selection tool may not provide sufficient justification for the use of the instrument in many hiring organizations. After all, a random-number generator does not require AI but can winnow down large applicant pools quickly without producing adverse impact. Despite the speed, scalability, and lack of bias of a random-number generator, it lacks the reliability, validity, and utility that hiring organizations expect in order to identify capable candidates and achieve an acceptable return on investment. It is incumbent upon developers (and subsequently users) to provide appropriate and sufficient evidence that AI selection tools meet these requirements. When AI-based recruitment and selection systems are rapid but non-random, other recruitment and selection issues may arise. A writer for *The New York Times* (Ajunwa, 2019) described a "closed-loop" system in which advertising attracted certain types of applicants who were automatically assessed, with those results used to focus future recruiting efforts. This kind of closed-loop system can sometimes worsen discrimination against job applicants, and the author makes the argument that plaintiffs should be allowed to bring suits against employers when they experience such discrimination, potentially leaving employers with the traditional legal burden of demonstrating the validity of the tools, even in the absence of adverse impact as usually assessed.

Organizations need employement tests that meet professional standards of fairness and accuracy; however, they must also be concerned about regulatory compliance. Of significant concern to I-O psychologists in the selection context is guidance from regulatory agencies, combined with statutory restrictions from state legislatures, that inform the evidence required to support the use of new technologically enhanced assessments. Since 1978, the *Uniform Guidelines* have specified the requirements for validity evidence when adverse impact exists. In July of 2019, the OFCCP published a new set of FAQs on employee selection procedures that were intended to clarify the *Uniform Guidelines*. In response to a question on selection procedures employing new technology such as "screening devices like games, challenges, and video submissions" that use artificial intelligence (AI) algorithms to assess qualifications, the OFFCP emphasized the historical requirement imposed on traditional selection measures that employers using such technologies provide validity evidence whenever adverse impact is found:

> Irrespective of the level of technical sophistication involved, OFCCP analyzes all selection devices for adverse impact. If OFCCP discovers that a contractor's use of an AI-based selection procedure is having an adverse impact at a contractor's establishment, the contractor will be required to validate the selection procedure using an appropriate validation strategy. (OFCCP, 2019)

We should take a moment to acknowledge that the *Uniform Guidelines* were adopted in 1978 and have had very few changes since then, and none since the adoption of the Q & As in 1981. (The OFCCP's 2019 FAQs were not adopted as actual components of the *Uniform Guidelines*.) Nonetheless, the *Guidelines* remain the controlling administrative rules governing much of the litigation involving Title VII of the Civil Rights Act of 1964 and is deeply embedded in case law for the last 40 years. Although both theoretically and practically out of date in some ways, they must still be considered in evaluating any selection procedure that results in adverse impact with respect to any protected group.

Several state legislatures have also weighed in on technologically enhanced assessments. One of the earliest to do so was the Illinois legislature. On May 29, 2019, the Illinois state legislature passed the Artificial Intelligence Video Review Act, which was signed into law by Governor Pritzker in August 2019 and took effect on January 1, 2020. This law requires employers who use video interview technology to evaluate job applicants' facial, speech, and other characteristics to notify each applicant in writing that AI and these characteristics may be used to evaluate fitness for the position, to provide a description of how the technology works and what characteristics are used, and to obtain written consent before the video interview. In addition, employers may not share the video with anyone other than those who have the expertise to evaluate it, and they must destroy the video within 30 days of the completion of the hiring process for the position. Of course, legislation and policies concerning the nature, transparency, and privacy of applicant data in technology enhanced assessments will continue to emerge and evolve.

Standards

Employment testing is not without important sources of professional guidance. Two documents guide research and practice in the area of employee selection and apply, regardless of the form of assessment:

- *Principles for the Validation and Use of Personnel Selection Procedures* (*Principles*, 2018)
- *Standards for Educational and Psychological Testing* (*Standards*, 2014)

In addition, in the U.S., the *Uniform Guidelines on Employee Selection Procedures* inform lawful employment testing. Regulatory agencies like the EEOC and the OFCCP provide supplementary guidance on employment testing through periodic questions and answers. (Some of this paper is decidedly U.S.-centric because of the legal requirements; however, many of the concerns discussed apply to test users globally.) As psychologists and members of SIOP, we also subscribe to the APA code of ethics (APA, Ethical Principles of Psychologists and Code of Conduct, 2010), including the treatment of candidates, and influences what we say about it.

Purpose

There are two purposes of this paper. First, we explore scientific, legal, and ethical concerns regarding new forms of employment testing, including those that are AI-based or technologically enhanced in some fashion. In the following sections, we discuss 11 concerns:

- Lack of a theoretical basis for predictors
- Job analysis
- Job relevancy
- Appropriate methodology
- Validity
- Reliability
- Changes to Technologically Enhanced Systems
- Control over the data presented to an employer
- Applicant experiences and reactions
- Communications
- Ethics

Some of these issues have been discussed previously in the context of big data analyses in general (Guzzo, et al., 2015); our focus is specifically on the selection context. For each topic, we have described the concern and then highlighted key questions that need to be answered. We have intentionally *not* provided our own detailed perspectives on these issues. Some of these questions may turn out to be more straightforward to address, whereas others are likely to be quite difficult and to require substantial discussion, input from other professionals, and probably additional research to address. Second, we argue that I-O psychologists must be central players in establishing how existing standards for employment testing should be applied to these new forms of testing. We close with a call to action for I-O psychologists working in the field of employment testing to address the concerns raised in this paper and establish professional standards for technology enhanced selection tools that are based on the *Principles*.

The comments that follow in the next section highlight concerns about new forms of employment testing. The order of presentation is not intended to imply relative importance but generally follows the flow of validation research, from job analysis, to predictor development, to criterion development, and so on. We have tried to illustrate most of the concerns with appropriate generic examples that reflect common practices that may apply to many different firms or individuals. Our intent is to be illustrative, general, and constructive in our approach—not to attack any particular test offering, method, or vendor.

Before we enumerate these concerns, a few points merit noting. First, not all concerns apply to every form of assessment. Second, some of the concerns are overlapping. For example, it is difficult to separate requirements for a job analysis from issues of job relevancy or the constructs and theories that underlie and define KSAOs and job characteristics. Third, many of the concerns presented here have existed for decades and are not unique to technologically enhanced forms of personnel testing. For instance, questions about what constitutes an appropriate job analysis methodology or about when a test needs to be revalidated have been long debated. What is new are the additional perspectives that new technologies contribute to the conversation, concerns, and even the solutions.

Concerns

Lack of a Theoretical Basis for Predictors

As indicated previously, many new technologically enhanced assessments use a wide variety of data that are obtained or "scraped" from applications, resumes, social media, emails, the Internet, or other sources; they are then evaluated using any of hundreds of possible machine learning algorithms. Although the choice of machine learning method is sometimes idiosyncratic to the researcher, the choice is often based on factors such as technical considerations, nature of the dataset, availability of software, and the researcher's familiarity with various methods.

The substantive nature of the included data, variables, and their linkages to job requirements are often unknown. As one example, data regarding past employers may be pulled from resumes or applications. One might find that former employment at Employers A, B, and C predicts future job performance at Company X, whereas employment with Employers D, E, and F does not—even though all six employers are in the same business, and there is no substantive *post hoc* explanation for the differences among the organizations. In another scenario, imagine that empirical findings indicate that educational coursework in theology predicts job success in sales—a job not ostensibly related to religion, philosophy, or charitable acts. Although such unusual relationships may in fact be driving the machine learning algorithm[3], they do little to support the work relatedness of the predictive relationship and build confidence in the selection system.

Other forms of technologically enhanced selection procedures are also atheoretical. For instance, no obvious theory or supportive data appear to be associated with how various aspects of an applicant's voice or facial characteristics relate to the KSAOs of an ideal job applicant (or job performance). Here, justification for prediction in terms of the characteristic's relationship to job requirements is inferred at best and unknown at worst. Similarly, a wide range of data related to characteristics of the test taker's responses are frequently collected (e.g., response time, changing answers). Although there is an apparent rationale for some measures (e.g., speed of response might be inferred to be a function of cognitive processing or cognitive ability), the theory and related research that might support that rationale usually go unstated, or assumed but untested.

I-O psychologists have long debated the need for a theoretical basis for predictors. Long before the use of big data and technologically enhanced tools, many I-O psychologists decried a lack of theoretical basis for many predictors used in personnel selection (e.g., assessment centers, biodata, situational judgment tests, interviews). Other I-O psychologists have taken roughly the opposite position: If scores on a set of predictors correlate with an organizationally relevant criterion (e.g., measures of job performance, engagement, or turnover), then those scores are useful predictors, and understanding the underlying rationale is considered as simply "nice to know."

In selection contexts, the theoretical basis or rationale for including measures within a personnel selection system is traditionally the extent to which each measure reflects a KSAO necessary to perform the job, as determined by a job analysis. I-O psychologists have focused on developing and analyzing theory-based and job-relevant psychological measures in order to rule out relationships that seem questionable (e.g., facial features and job performance) or biased (e.g., race/ethnicity covariates and job performance). By contrast, when an algorithm is applied to data scraped from various sources, there are rarely theoretical underpinnings to the choice of predictors (Braun & Kuljanin, 2015), which may be massive (big data), messy (multiple data sources), and missing. The algorithms can be difficult to interpret, and when they are interpretable, the relationships discovered may have little obvious practical or conceptual relevance to the work being performed. Instead, the strength of the relationship between the predictors (or 'features' in data science terminology) and some criterion becomes suffiicient justification for the use of the selection procedure.

In the absence of adverse impact, even job-irrelevant predictors would not require evidence of being "job related and business necessity" (Title VII, Civil Rights Act of 1964) or be unlawful *per se*. From this perspective, the

[3] Note that the variables actually driving predictive relationships are frequently not easily discoverable; however, when they are, end users often note such unusual relationships.

fact that selection decisions are based on big data that are convenient (web scraping) or fun (games) or unrelated to required KSAOs is immaterial. In this context, explanation and theory might be viewed as something that I-O psychologists merely like to think about; they are intellectual exercises that can be easily avoided so long as there is no adverse impact. Yet, the *Standards* and the *Principles* emphasize the importance of a theoretical basis for selection procedures in their shared definition of validity: "the degree to which accumulated evidence and theory support *specific interpretations* of scores from a selection procedure entailed by the proposed uses of that selection procedure" (Principles, p. 96; Standards, p. 225; emphasis added). If the *relationship* between the predictor and the criterion supports the intended *interpretation* of the score, then that relationship might be interpreted as evidence of validity and justify the use of the test.

If the only purpose of a selection tool were mechanical, that is, to predict scores on a measure of job performance (or another criterion), then investigation of the underlying constructs of the predictors and substantive study of jobs and their requirements would be merely a response to regulatory requirements. However, if the purposes look beyond simple prediction, then understanding the predictive relationship can lead to improved assessment measures, increased coverage of the performance domain, greater generalizability, and assurance that selection systems are sensible in terms of recruiting, organizational training efforts, diverse applicant pools, and changes over time. Understanding work and its requirements necessitates both scientific research and practical thinking, which go beyond data that are conveniently obtained or algorithms that mine complexity in the data for nonobvious relationships (Rotolo & Church, 2015). Systematic research enables the identification of additional variables and data sources that may also predict, mediate, or explain work behaviors.

The underlying issue here devolves into a question of whether selection research is propelled by science, with a premium on understanding applicants' suitability through the lens of job requirements, or whether selection research can be atheoretical, solely an empirical activity intended to maximize predicted outcomes.

Questions:

- *Are theoretical justifications necessary in employment testing?*
- *Is a technologically enhanced selection measure that predicts organizational outcomes sufficient, or does one need to understand why that prediction occurs?*
- *Do theoretical justifications improve practice in employment testing?*
- *Do the considerations about the theoretical justification of selection procedures change when there is adverse impact versus when there is not?*

Job Analysis

Professional standards, legal guidelines, and case law emphasize the need for selection systems to be tied to job requirements, which are often determined through an analysis of the characteristics of the worker and the work to be performed (Morgeson et al., 2020). Traditionally, job analysis forms the basis for the identification of the KSAOs required for job performance and the appropriate variables to consider for selection into jobs and for the development of relevant criteria for assessing their validity, such as defining the domain of job performance.

The *Uniform Guidelines* highlights the importance of reviewing job information for criterion-related validity studies:

> *(2) Analysis of the job.*
> There should be a review of job information to determine measures of work behavior(s) or performance that are relevant to the job or group of jobs in question. These measures or criteria are relevant to the extent that they represent critical or important job duties, work behaviors or work outcomes as developed from the review of job information. (Uniform Guidelines, Section 14 B (2))

The *Principles* also makes clear that the purpose of the job analysis is to define appropriate predictors and establish the relevancy of the criterion measures used.

> In the context of validation research, there are generally two major purposes for conducting an analysis of work. One purpose is to develop or identify selection procedures. Part of this development process is an analysis of work that identifies worker requirements, including a description of the KSAOs or competencies needed. Such an analysis would determine the characteristics workers need to be successful in a specific work setting or the degree to which the work requirements are similar to the requirements for work performed elsewhere. The second purpose is to develop or identify criterion measures by assembling the information needed to understand the work performed, the setting in which the work is accomplished, and the organization's goals. (Principles, p. 12)

Importantly, neither the *Uniform Guidelines* nor the *Principles* specifies a particular method of job analysis; instead, they acknowledge there are many different acceptable ways to identify job requirements. Also, note that some exceptions to the expectation of a job analysis are allowable, for example, when demonstrating the importance of certain criteria, such as turnover or counterproductive work behaviors (CWBs, such as stealing), because such criteria are considered clearly relevant to all jobs.

In practice, many organizational test users, both those using traditional forms of tests and those using technologically enhanced forms, fail to conduct job analyses at all, or they conduct an abbreviated form of job analysis. Some employers eschew job analysis altogether, when they hire for a particular position and deploy a selection procedure that simply purports to measure KSAOs that are (in their judgment) obviously required for the job in question. For example, they may choose a "sales test" to select candidates for a sales job. For myriad reasons, employers may assume that all sales positions are alike, regardless of the company, and the transference of job analytic and validation efforts in other organizations are therefore relevant, without additional effort. However, without the organization conducting a job analysis, it will not be clear the extent to which (a) sales skills purportedly measured by the test are the skills that predict sales success in a specific organization, and (b) the particular sales skills required by the employer's position are the same as those measured by the test.

Necessity of job analysis when the predictors and criteria are strongly related in a criterion-related study. Closely related to the question of the need for theoretical underpinnings is the question about the need for a job analysis that justifies the inclusion of predictors and criteria in a selection system. If a set of predictor variables appears to be job relevant and has been found to predict a criterion of interest, does it matter whether its relationship to the requirements of the job in the local setting is established through a credible job analysis? If job analysis was not used to justify a criterion measure already in use, does it still need be conducted, or is the fact that the organization uses the criterion for another purpose (e.g., to assess performance) provide sufficient justification?

Some argue that a well-established empirical relationship between predictors and criteria (either from local validation studies or meta-analytic studies) alone indicates job relatedness. However, unless one considers additional substance and context, there are serious risks associated with this perspective. Consider the situation in which past "leadership experience" is generally found to be positively correlated with job performance in a managerial position; yet, further validation research finds that gender also correlates with job performance, such that men tend to have higher performance ratings. Although gender should be irrelevant to managerial performance, the relationship between gender and performance might be found because men tend to be given and accept more leadership opportunities. In the financial context of lending practices, a recent *Wall Street Journal* article reported a similar situation with race:

> It's well known, for instance, that in credit scoring, ZIP Codes can serve as a proxy for race. AI, which uses millions of correlations in making its predictions, can often base decisions on all sorts of hidden relationships in the data. (Totty, 2020)

Without a job analysis to identify, define, and measure the KSAOs relevant to managerial job performance, predictors and criteria may be used that are wrong, inadequate, or unfair. (This issue is discussed further in the section "Job Relevancy" below.)

Forms of job analysis. If the question of the need for a job analysis is answered in the affirmative, then there is the more complex question of what type of job analysis is to be conducted. The legal and professional guidelines for employment testing are clear that there are many different, acceptable approaches to analyzing work; however, they are silent on the question of how comprehensive the job analysis must be. Although job analysis is most often a central issue in cases involving content validation strategies (cf., Guardians Association of New York City Police Department, Inc. v. Civil Service Commission of City of New York, 630 F.2d 79 (2d Cir. 1980)), legal challenges to employment tests, in general, have suggested that the job analysis should be systematic and accurate, regardless of the methodology used.

Many publishers of both traditional selection tests and those based on technologically enhanced systems use a standard job analysis process across organizations that is based on some form of competency model, with no close ties to actual tasks or work behaviors performed by incumbents in the job. These competencies may reflect relatively specific KSAOs (e.g., written communication skills) and/or broader forms of KSAOs (e.g., general communication skills). The competencies may also be broad work behaviors, stated in some cases as organizational aspirations (e.g., "Demonstrates amazing customer service at all times") rather than KSAOs. Most competency models offered by test vendors comprise competencies that are relevant to a wide array of jobs (e.g., Hunt, 1996, outlines eight general KSAOs required for "generic" work performance across a large set of entry-level jobs, such as industriousness, attendance, and theft). Organizations then rely on the same competency list to fulfill the job analysis requirements for all jobs and job levels, with great efficiency.

There is no assurance that a generic competency list will be complete relative to the requirements of a specific job. But how complete does the list of competencies need to be? Many test publishers have competency or KSAO lists that are related to their tests. If the test offering does not purport to measure the KSAO, it (the KSAO) is frequently not included on the list. Although there is no requirement to measure every critical competency for a job, prediction of a broad criterion like job performance is often enhanced by including measures of the most critical competencies.

Whatever form of job analysis is chosen, and by whatever means job analysis is accomplished, I-O psychologists are aware of and must work within the practical demands (time, money, staff) that job analysis places on organizations. They frequently observe the reluctance of an organization to undertake a comprehensive analysis of a job for the purposes of test and criterion development and validation. The reasons for abbreviated approaches to job analysis are abundantly clear to every practitioner who has had to convince a line manager of the importance of a job analysis: The shorter, the less intrusive, and the less expensive, the better. Furthermore, convincing managers of the advisability of a job analysis, particularly when the managers believe they know the job well, is a daunting task. Any process that reduces the demands placed on managers, supervisors, and employees and makes such an analysis effort more feasible will help persuade participation.

Rigor in the job analysis methodology to establish job requirements. If some form of job analysis is necessary to establish job relevancy, then we must return to the questions about appropriate methodologies for job analysis, including those regarding the number of SMEs needed to accurately define and reliably rate tasks and KSAOs. How much direct contact (e.g., observation, interviews), if any, is necessary for an analyst to sufficiently understand a job? To what extent do job complexity and unobservable work activities or behaviors affect the type and level of job analysis required? What level of job analysis is necessary to establish the relevance, breadth, and acceptability of the criterion measures used in any analysis of potential predictors, whether using traditional analytical tools applied to job analysis ratings and/or AI algorithms applied to new forms of big data? I-O psychology has recommendations to offer in response to these questions but rarely has bright lines that define a sufficient job analysis. Professional judgment and decision making will always be required, accounting for the critical contextual factors and goals of the job analysis.

Using the O*NET. Some organizations have relied on the worker and work characteristics contained within the O*NET occupational database, primarily the knowledge and skills sections, but sometimes the tasks and the work activities sections. Because of the manner in which the O*NET data were collected, they may be more or less relevant to a particular job in a particular organization. Thus, many suggest that O*NET data are useful as a starting point, but they should not serve as a complete job analysis.

> O*NET data (http://online.onetcenter.org/) should *not* [emphasis in original] be used as a substitute for conducting a formal job analysis given that it possesses certain limitations... However, O*NET is an extraordinarily useful resource for job analysts wanting to acquire information about jobs and aids in the development of preliminary job tasks and KSAs before conducting a systematic job analysis. (Gutman, et al. 2011, pp. 191-192)

Collecting information on tasks (e.g., frequency and importance ratings). The *Uniform Guidelines* requires measures (usually in the form of ratings by incumbents or SMEs) of task importance or criticality, which is often operationalized as some combined form of frequency and importance.

> A description of the procedure used to analyze the job or group of jobs, or to review the job information should be provided (Essential). Where a review of job information results in criteria which may be used without a full job analysis (see section 14B[3]), the basis for the selection of these criteria should be reported (Essential). Where a job analysis is required a complete description of the work behavior(s) or work outcome(s), and measures of their criticality or importance should be provided (Essential). The report should describe the basis on which the behavior(s) or outcome(s) were determined to be critical or important, such as the proportion of time spent on the respective behaviors, their level of difficulty, their frequency of performance, the consequences of error, or other appropriate factors (Essential). Where two or more jobs are grouped for a validity study, the information called for in this subsection should be provided for each of the jobs, and the justification for the grouping (see section 14B[1]) should be provided (Essential). (*UGESP*, Section 15B(3))

This step of collecting ratings on both tasks and KSAOs is often skipped. Typically, the reason offered for collecting rating data on tasks is to orient SMEs to actual job requirements and steer them away from stereotypical beliefs. Although sparse, some research data support this contention. Morgeson et al. (2020) reached this conclusion after reviewing the literature:

> Although there is little in the way of empirical data on the issue, we suspect that reliability and validity of judgments about KSAOs will increase to the extent that the KSAOs are relatively concrete and closely tied to specific tasks. (p. 375)

What subject matter experts (SMEs) are asked to do with competency lists in the job analysis context often varies. Sometimes, SMEs rate the importance of the competencies or the extent to which they are needed at entry; other times, they are asked to identify some number of competencies from a broader set that are most important to the job. The accuracy of SMEs' judgments about the competencies for a job can be assessed through interrater agreement (intraclass correlations), but this assumes they are similarly knowledgeable or trained. Neither the assumption of similar knowledge or training nor the level of interrater agreement is regularly evaluated in most practice.

According to professional standards like the *Principles* and the *Standards*, some form of job analysis is best practice, even though the minimally acceptable form of that job analysis goes unspecified because it is conditional on the purpose at hand, the nature of the job, the criterion of interest, the nature of the selection instrument, and the type of validation study to be undertaken. Nevertheless, job analysis is often viewed by organizations as an unnecessary activity, more of a bureaucratic burden than of real value. Although the demands of any job analysis can be onerous, they are particularly unpalatable when client organizations believe the requirements of the job are obvious. Although best practice in I-O psychology promotes job analysis to identify job requirements empirically and establish the job relevancy of predictors and criteria, line managers may not agree or be willing to devote time and resources to the endeavor.

***Questions**:*

- *Is it necessary to conduct a job analysis if the predictors and criteria are strongly related in a criterion-related study?*
- *Is a job analysis necessary to justify an operational performance measure (e.g., key performance indicators) that will be used in the validation study?*
- *To what extent is a competency model an adequate substitute for a job analysis?*
- *Is it important to have a complete list of competencies?*
- *How much rigor in the job analysis methodology is necessary to establish job requirements?*
- *Is the O*NET an acceptable source for a complete list of KSAOs?*
- *Is it important to collect information on tasks (e.g., frequency and importance ratings)?*

Job Relevancy

Closely related to concerns about the theoretical basis for a test and the need for a job analysis is the issue of job relevancy. *Job relevancy* refers to the core features of a job as determined through a job analysis (e.g., KSAOs, work behaviors, and environmental characteristics), and we will use this term interchangeably with the term *job relatedness* (although the perspective of Guion, 2011, reflects distinctions between these concepts worth appreciating). U.S. federal laws require job relatedness for tests used for employment decisions when adverse impact is observed. Title VII of the Civil Rights Act of 1964 states:

> An unlawful employment practice based on disparate impact is established … only if a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin and the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity. (Sec. 2000e-2 (k)(1)(A)(i))

In establishing validation requirements under Title VII, the *Uniform Guidelines* uses the term *job relatedness* in the context of ensuring that selection measures are tied to requirements of the job and establishes the explicit requirement to conduct job analysis whenever undertaking a criterion-related validation study (*Uniform Guidelines*, 3.14(A)).

Unlike the *Uniform Guidelines*, the *Principles* highlights several empirical approaches to establishing job relatedness in its definition:

> The inference that scores on a selection procedure are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the job relevance of the content of the selection instrument, or of the construct measured. (*Principles*, 2018, p. 90)

The *Principles* adopts the same definition of validity adopted by the *Standards* as "the degree to which *evidence and theory* [emphasis added] support the interpretations of test scores for proposed uses of tests" (*Standards*, 2014, p. 11). The *Standards* is clear in stating that validation includes "an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure" (*Standards*, 2014, p. 11).

The *Principles* and the *Uniform Guidelines* both emphasize the importance of job analysis as part of the validation process. Job analysis is essentially the pursuit of job relevance, a necessary requirement before one can interpret validity evidence for job relatedness. Throughout both the *Principles* and the *Standards*, we see the presumption that validity is a function of evidence supporting interpretation of assessment outcomes (e.g., test scores) for specific purposes. As we have previously noted here, technologically enhanced selection tools may not incorporate predictors or "features" that are demonstrably job relevant via job analysis, although they may be job related in the sense that they are predictive of outcomes, yet in turn may themselves not be based in job analysis.

Questions:

- *If a job analysis is not conducted, can job relevancy be demonstrated for either the proposed assessments or the criteria against which they are evaluated?*
- *Is the correlation between predictor and criterion alone sufficient to establish job relatedness under the* Uniform Guidelines*?*

Appropriate Methodology

Technologically enhanced selection processes may apply machine learning algorithms to thousands of data points, weighting and combining the data in ways that attempt to make predictions that (a) may be highly complex in nature (interactive, nonlinear) and (b) hold up in new data sets independent of the one on which the prediction was developed. Because of the nature of many machine learning algorithms, results may be obscure, counterintuitive, and otherwise difficult or impossible to interpret (e.g., random forests base predictions on hundreds of "trees"; neural nets tune arbitrary and layered configurations of "neurons"). Overall, understanding what is going on inside the "black box" of machine learning can be problematic. Although meaningful strides continue to be made in explainable AI (XAI) within many areas of machine learning (Ribeiro et al., 2016) and computer vision (Kaleghi, 2019), we are unaware of any breakthroughs in understanding complex prediction in selection that would yield new insights for theory or practice.

When these AI methods are used in talent assessment, the underlying predictive structures are generally inaccessible or proprietary. Moreover, machine learning methodologies may be unfamiliar, if not completely foreign, to many I-O psychologists in both practice and research, making it difficult to evaluate the results and interpret them for ourselves, our stakeholders in organizations, the legal community, and beyond. That said, many I-O psychologists have strong training in psychological measurement and psychometrics. If they make the effort to learn the fundamentals of machine learning, this extension of their knowledge can enable them to participate in very important conversations, such as whether a big data analysis was necessary, whether it provided meaningful prediction (and a substantial improvement over other methods), and whether and when the model predictions are generalizable to other samples. Much more machine learning education of I-O psychologists is clearly needed because its use is becoming more widespread in talent analytics (for some guidance, see Aiken & Hanges, 2015; Oswald & Putka, 2016, 2017).

Criteria for appropriate technologically enhanced models. AI-based algorithms generate a variety of metrics to evaluate the final selection model derived, including concepts such as minimizing mean-squared error; understanding the confusion matrix that compares actual with predicted values; and the receiving operator curve (ROC), which illustrates correct and incorrect classifications under different cut scores as summarized by area under the curve (AUC). Again, many I-O psychologists lack the background to interpret and evaluate such metrics; yet, they have enough methodological training to benefit from the additional education needed to understand these measures.

One issue that deserves more attention is the high ratio of variables to sample size often used in big data applications and machine learning-based algorithms. In big data applications, this ratio might be 30 variables per case, for example, whereas in traditional analyses, it might be the opposite, one variable for every 30 cases. Traditional statistics are literally impossible in the big data situation where there are more variables than cases (e.g., in linear regression analysis, the variance–covariance matrix of predictors will not invert), which means that machine learning is a necessary analysis tool if one decides to operate on big data (and not reduce the set of variables via composites informed by factor analysis, scale scores, etc.). Even though machine learning algorithms almost universally incorporate cross-validation to assure robust prediction, interpretation of results differs from traditional variable-based approaches (e.g., regression coefficients). The variable-driven interpretation of algorithmic predictions when the number of variables exceeds the number of cases, often remains in a black box, and the continued development of XAI in the context of personnel selection will be extremely important to achieving transparency.

Finally, the available literature on the application and efficacy of these machine learning algorithms in personnel selection contexts is not sufficient to provide a basis for making informed comparisons. For more traditional forms of testing, we have a large meta-analytic research literature that provides typical ranges of correlation coefficients for different types of selection instruments measuring different constructs (Schmidt & Hunter, 1998), and most I-O psychologists are familiar with Cohen's rules of thumb for effect sizes (Cohen, 1988). In addition, effect-size benchmarks have been established in substantive domains (Bosco et al., 2015). Given our empirical knowledge of traditional measures and methodologies, I-O psychologists would be highly skeptical of a .75 correlation between a score on a structured interview and a measure of overall job performance. But what is the strength of the predictive relationships to be expected from machine learning algorithms applied to data from new technologies? Many I-O psychologists currently lack a fundamental understanding of how different machine learning algorithms work, their assumptions and other boundary conditions, and the metrics they produce, thus making it challenging to compare the results of AI based algorithms to one another and to the results of traditional multiple regression analysis.

Questions:

- *What are the appropriate criteria for technologically enhanced models?*
- *What is the best approach to choosing and using a machine learning algorithm?*
- *What defines acceptable results when machine learning algorithms are used? What would be an acceptable level of prediction?*
- *How can we determine if the predictive results are based on idiosyncrasies of the sample on which the model was built? How generalizable are the results to other samples?*
- *What continuing education experiences and changes to graduate education in I-O will be needed to prepare I-O psychologists to develop, research, and evaluate technologically enhanced selection tools?*

Validity

Among I-O psychologists, there is no doubt of the importance of validity in employment testing. Evidence of validity for the intended inference is necessary to demonstrate the effectiveness of the selection procedure and is a legal requirement when adverse impact is observed. For business reasons and legal defensibility reasons, validity is a *sine qua non* of selection research.

When AI-based predictors, game-oriented selection tools, and evaluations of facial features and voice qualities are used to determine whom to hire, they and similar methods are considered tests both under the law and according to professional guidelines. Section 2B of the *Uniform Guidelines* clearly indicates that a test is any selection procedure used for an employment decision, which connotes a broad array of organizational decisions about an individual.

> *B. Employment Decisions.*
> These guidelines apply to tests and other selection procedures which are used as a basis for any employment decision. Employment decisions include but are not limited to hiring, promotion, demotion, membership (for example, in a labor organization), referral, retention, and licensing and certification, to the extent that licensing and certification may be covered by Federal equal employment opportunity law. Other selection decisions, such as selection for training or transfer, may also be considered employment decisions if they lead to any of the decisions listed above. (*Uniform Guidelines*, 2B)

Consequently, in the U.S., the use of these tests must be supported by evidence of validity if adverse impact results from their use. In all countries, including the U.S., validity evidence provides a business justification for the use of the selection system.

Evidence of validity has typically been evaluated in employemnt settings using either criterion-related validation or content-oriented validation strategies. According to professional guidelines, evidence derived from criterion-related validity studies is commonly demonstrated by establishing a statistical relationship between employment test scores and some relevant criterion such as job performance, most commonly by using correlation and regression techniques.

> Evidence for criterion-related validity typically consists of a demonstration of a relationship between the scores on a selection procedure (predictor) and one or more measures of work-relevant behavior or work outcome (criteria). (*Principles*, 2018, p.14)

The strength of a linear relationship between predictor and criterion is usually evaluated by the size, confidence interval, and statistical significance of the correlation coefficient. Moreover, regression models should provide some indication of imprecision in their estimates to evaluate how well they predict (e.g., confidence intervals, bootstrapping, cross validation). Nonlinearities are rarely modeled in the traditional selection context because linearity affords greater interpretability (Coward & Sackett, 1990; Walmsley et al., 2018), and nonlinearity tends to require much greater statistical power (Converse & Oswald, 2014). In contrast, machine learning algorithms attempt to model and cross-validate nonlinearities, and even with a large sample size, interpretation is difficult at best. Statistical power may or may not be at issue depending on the characteristics of the dataset (both the number of variables and the number of cases). Because of the nature of AI-based tests and machine learning algorithms, content validation evidence (where SMEs relate the content of the test to the requirements of the job) is often not used at all or used only to supplement a criterion-related validation study.

Some machine learning algorithms are based on linear regression analysis (e.g., logistic, lasso, ridge, and elastic net regression), and thus, they more readily provide coefficients that can provide a "demonstration of a relationship" as noted in the *Principles* above. Other models use more complex prediction models (e.g., random forest, neural nets), or they take a clustering approach to selection (e.g., a match to the scores of "good" performers in the organization). In these latter cases, relationships with the criteria are not necessarily as apparent or direct and not always expressed in terms of a coefficient. Overall model fit (analogous to the R^2 in regression analysis) might be the only basis for establishing the predictor–criterion relationship, perhaps making job analysis all the more important to at least ensure that job-relevant predictors and criteria are being used in the machine learning algorithm, even if it is not clear how those variables are weighted to produce predicted values.

Fundamental requirements for validity. An important question is whether users of selection procedures based on algorithms should be required to provide specific conceptual and empirical evidence of predictor–criterion relationships, in addition to overall metrics used to evaluate the predictive power of machine learning algorithms (see job analysis discussion above). Because employers in the U.S. are obligated to search for alternative selection procedures that have equal or greater validity and less adverse impact, comparative data are particularly important. (See Section 3B of the *Uniform Guidelines*.) Empirically, correlational results and our history of meta-analytic findings in I-O psychology might be established as a reasonable baseline for machine learning algorithms to beat.

Minimum requirements for documentation. To varying degrees, the *Uniform Guidelines*, the *Principles*, and the *Standards* all enumerate the important components that should be documented in technical reports for personnel selection procedures. Yet, critical information is often omitted from technical reports that would allow for better evaluation of the procedures and evidence supporting them. The problem is particularly acute with technologically enhanced assessments, where a large number of decisions may be made in the process of selecting and tuning a machine learning algorithm and then applying its results. Important questions to be addressed are both general and specific in evaluating this research. For instance, how were data obtained, prepared, cleaned, transformed, and combined? What approach(es) were used to address missing data? With text mining of resumes, personal statements, or other written responses, how are effects due to vocabulary knowledge or verbal fluency controlled, in cases where the text is not intended to measure the verbal ability of the job applicant? How are conflicting applicant data in a dataset weighted, discarded, or cleaned? Was only one algorithmic approach taken, or several, and could an external party reliably reproduce the machine learning analysis? Were there *a priori* decisions made about which algorithm(s) to use? Were predictions combined across algorithms? Finally, was there any *post hoc* filtering (i.e., inspecting the results to select the algorithm that looked good, which is a suspect practice)?

There is a clear need for transparency on the part of those who develop tests using innovative methodologies. They must be forthcoming with information that allows others to fully understand and evaluate their work and for users to comply with federal guidelines and requirements. Yet, even when data scientists are transparent in communicating what they do, I-O psychologists and other users and reviewers may still be confronted with complex interactions and nonintuitive relationships that cannot be easily expressed as a mathematical function otherwise interpreted until more “explainable AI” tools are available.

Questions:

- *What is sufficient evidence of validity when machine learning models are used?*
- *What details of AI research must be documented?*
- *What details of AI research must be shared with users?*

Reliability

Like validity, reliability is another absolute requirement for any test. The *Uniform Guidelines* requires the reliability of selection procedures to be evaluated and reported (for example, see Section 14C(5), Section 15B(7), Section 15B(8)). Professional guidelines also emphasize the importance of the reliability of the predictors:

> ***Predictor reliability.*** The scores obtained from predictor measures should exhibit adequate levels of reliability. The factors critical to addressing the issue of reliability of criterion measures that were discussed earlier apply to predictor measures as well (e.g., identifying the conditions of measurement across which one wishes to generalize the scores of interest, adopting a study design that will allow for calculation of reliability estimate(s) that evaluate whether scores generalize the said conditions). Once again, in the event it is not possible to gather such data as part of the predictor development or criterion-related validation effort, results regarding the reliability of predictor scores should be qualified accordingly. (*Principles*, p. 22)

To be reliable, scores from any assessment must be measuring KSAOs that are relatively stable and generally consistent over time and setting. For example, job applicants’ scores should be about the same if they were to take the test again a week later (memorization notwithstanding). Additionally, a traditional expectation is that irrelevant situational factors or inherent inconsistencies in an assessment do not materially affect the observed scores. Selection tests cannot be contingent on factors such as day or time of testing, variations in testing conditions, or the particular equipment used, unless such variations can be shown to be job relevant. Reliable changes in test scores should be due to changes in the individual’s standing on job-relevant constructs (e.g., increases in job-relevant skills) or to decreases in sources of irrelevancies (e.g., less anxiety, greater understanding of the test protocol). Similarly, changes in scores on a game should not be due to the idiosyncrasies of the game or changes in baseline skills (e.g., multikey or mouse use beyond what is normally required).

However, reliability evidence for technologically enhanced selection measures is often minimal or difficult to obtain, but when available, the results are mixed. On one hand, some evidence suggests that the reliability of machine-scored employment interviews is high and, in some cases, higher than for interviewer-scored interviews. Interviews scored by machines typically have substantial agreement with human scorers when they are well trained. In practice, however, human scorers of interviews in operational settings are notoriously prone to error and often demonstrate low interrater agreement. There has been long-recognized potential for greater reliability and consistency in algorithmic scoring (Kuncel et al., 2013), which speaks to the potential for greater fairness—so long as job-relevant information is being scored.

On the other hand, reliability evidence for the emotions extracted from facial feature analysis, even if they were actually relevant for employment, is particularly mixed. One study (Cowen & Keltner, 2019) involved the human judgment of photographs representing different emotions and concluded that facial expressions can reliably signal at least 28 different categories of emotions. Other studies suggest that facial features of people with darker skin are more difficult to evaluate via artificial intelligence than those of lighter skinned individuals (Singer & Metz, 2019). Although not yet well researched, it is not clear how the facial features of people who have injuries and disabilities that alter facial features, who take medication that changes their appearance, or who have altered facial features (e.g., scars, tattoos) would be treated.

Threats to reliability may also come from individual differences in the data that are collected rather than the manner in which they are measured. For example, relatively more extraverted applicants may provide greater detail in written or spoken information compared to less extraverted ones and, in doing so, provide the algorithm with more key words that may be related to other traits such as cognitive ability or verbal skills, which are not necessarily related to extraversion. Consequently, whether it is due to big data, machine learning algorithms, or both, one trait may end up affecting the reliability of other traits.

Questions:

- *How should reliability be assessed when AI is used to build predictive models used for selection?*
- *What should the minimal requirements for documentation of reliability be?*
- *Are appropriate and sufficient measures of reliability available and reported for technologically enhanced assessments?*
- *Are the new forms of employment assessment sufficiently reliable to meet psychometric standards?*

Changes to Technologically Enhanced Systems

Dynamic models and norms. Innovative selection procedures that use models derived from machine learning algorithms present opportunities for data analysis that were not feasible in the past. Many vendors advocate refreshing the algorithms frequently, sometimes even after every test administration. Often, these selection procedures are characterized as "dynamic," indicating that validation evidence and normative data are updated in near real time. Yet, such practices highlight a potential dilemma. On one hand, there could be real change in the nature of the applicant sample and/or the job, necessitating a change in the algorithm. On the other hand, changes in the algorithm may reflect sample idiosyncrasies or other instabilities over time that should not be capitalized on.

Whether this "dynamic" feature is useful remains to be seen for several reasons. First, there is no doubt that evidence of validity must be documented in a technical report. Thus, updates of underlying selection processes require updates to technical reports documenting the validity of the selection procedures as well as the characteristics of the normative data base. Out-of-date reports could be problematic for several reasons, including legal defensibility and HR records maintenance over time.

Second, each new resulting selection process potentially requires score adjustments to assessment results already in the data base or alternatively a policy on how to treat scores based on different processes derived at different times. The option of continually adjusting test scores may create significant administration problems. Candidates who are qualified today may become unqualified tomorrow (or vice versa) on the basis of changes in the selection process. Note that such changes could include altering the weighting of predictors, adding or removing predictors, and changing interpretations, including adjusting predictor cutoff scores or other means of score interpretation. Policies for treatment of past scores (e.g., various forms of grandparenting) may also be administratively difficult to manage in large-volume testing programs. One result of such changes may be that different candidates are evaluated on different variables depending on when they applied, raising the specter of disparate impact if a relationship between evaluation method and a proscribed group characteristic should result. Even in the absence of group-level disparate impact, concerns about disparate treatment could arise. Additionally, the appearance of such treatment might result in applicant dissatisfaction and public complaints.

Third, large shifts in the applicant pool may also have a major effect in terms of the reliability, validity, range restriction, or range enhancement on big data, so there is a need to monitor the key characteristics of the applicant pool (e.g., demographic, educational). Also, shifts in the technologies, data, and algorithms available in the future might change the nature of the applicant database being analyzed, which may indicate the machine learning algorithms used should be updated.

Fourth, defining appropriate group comparisons becomes much more difficult. For example, defining relevant applicant pools for analysis of adverse impact or defining appropriate normative groups for comparison would be challenging if an algorithm changed frequently.

Revalidation and norms updating. It is important to remember that employers do revalidate tests and update their norms. The difference between past practices and current ones related to the AI-based algorithms seems to be the frequency with which it is done. Traditionally, tests were validated again when there was reason to believe that the job had changed, the test had been compromised in some way, the characteristics of the applicant pool had substantially changed, or the time since the last validation effort was so long as to bring into question the usefulness of the validity evidence in a legal challenge. Revalidation was generally undertaken at well-spaced intervals because it was such a laborious task. With the computing power available today, this continuous updating is much less laborious, but doing so raises the issue of how often validation should be refreshed to accommodate the nature of new applicant data.

Questions:

- *Are dynamic algorithms and norms useful?*
- *How should results from dynamic algorithms be documented to comply with existing and future legal and administrative requirements?*
- *How frequently should tests be revalidated and norms updated?*
- *What are the indicators that revalidation and updates to norms are needed?*

Control Over the Data Presented to an Employer

Traditionally, applicants have been able to choose to a large extent what information about themselves to present to an employer. In general, applicants can "put their best foot forward" and manage the image they present to potential employers. They control the extent of their effort on ability, skill, and knowledge tests, the answers they choose to personality or situational judgment measures, their answers and demeanor in interviews, and the content of resumes and application forms. Of course, use of information outside of the resume, application, or test results is not new. "Word of mouth" has been a factor in selection for a long time, and employers have frequently consulted references and conducted criminal background investigations or credit checks for sensitive jobs.

More recently, however, employers have gained access to a wider array of information, some of which is decidedly not within applicants' control. Employers may search the Internet or large databases for information, especially for evidence of inappropriate behavior that might signal poor judgment or other undesirable characteristics. However, such evidence may also contain irrelevant information, such as demographics (Zhang et al. 2020) and political affiliation (Roth et al., 2020). Applicants generally have increasingly less control over the type and relevance of personal data that organizations can extract from social media and other sources on the Internet. Applicants may seek to engage in impression management through their social media profiles (Schroeder & Cavanaugh, 2018); however, in some cases, applicants may not have posted the information themselves, or the information posted has been substantially altered. In other situations, they may have done so without intending for it to be widely shared. Online information now available to employers may be highly suspect, substantially dated, or without context. There are companies that help individuals manage their online reputations, but they are not without cost. Moreover, to the extent that the knowledge of these services and their affordability varies by race/ethnicity or other demographic variables, these "scrubbing" services may contribute to adverse impact, which may be difficult to detect.

Through technology, employers now have access to an array of information that bears the promise of being predictive yet may be questionable in terms of job relevance, for example, images, video recordings, audio recordings, and measurements of autonomic responses such as facial micro-expressions and voice analysis that purport to convey job-relevant emotions. Most obviously, physical appearance, beyond simple grooming, is outside the control of most people. We cannot easily select or alter the color of our skin, the timbre of our voices, our basic speech patterns, or the characteristics of our faces. The use of these types of data raises serious ethical issues due to their job irrelevance. Furthermore, as noted earlier, facial appearance and vocal features may present special problems for people who look or speak differently due to cultural differences (e.g., minorities, immigrants), physical differences (e.g., disabilities, diseases, or injuries), gender differences, and age differences.

The point is that irrelevant variables may well predict job performance; here, the essential question is "Is it fair?" There is no law or guideline that says an applicant should have control over what the employer sees (except perhaps in the realm of privacy statutes). Nor is there a specific ethical standard that requires an employer to use only the data presented by the candidate. Yet, there seems to be a moral dilemma regarding what is fair game in the selection process. There appear to be several approaches to this conundrum. At one extreme, the answer is simply to press ahead and use this kind of data in selection with the rationale that historically applicants have never had control over everything an employer sees and uses for selection. At the other extreme, the solution is to use no data that are beyond the control of the candidate, which might entail careful review of traditional forms of assessment, such as biodata. A more moderate approach, which is being legislated in laws like Illinois' Artificial Intelligence Video Act, is to require informed consent before an employer bases a selection decision on data beyond the applicant's control.

Questions:

- *Is it fair to use data that are outside the control of an applicant?*
- *Should employers seek out data on the Internet?*
- *Would there be legal issues associated with not seeking information about some kinds of behaviors (e.g., poor judgment, behavioral deviancy, CWBs)?*
- *How long should applicants' past failures or mistakes affect their future job prospects and what mistakes should be considered (e.g., criminal history, online behavior, early life behavior)?*

Applicant Experience and Reactions

Many employers are concerned about attracting applicants—particularly highly qualified ones—when the labor market is relatively strong. Consequently, they strive to make the selection and hiring processes simple, quick, and engaging. Although salience of the candidate experience may vary with the unemployment rate, employers often remain concerned about a negative candidate experience discouraging highly qualified individuals, even when there is a plethora of job applicants. Many technologically enhanced assessments are particularly useful with respect to these goals because they require little effort on the part of applicants, or they are highly engaging. At the same time, some innovative methods of testing raise concerns about applicant experiences and reactions, leaving questions like the following unanswered for applicants or employers:

- How should applicants prepare for the assessment (are practice sessions allowable)?
- What is being measured, is it relevant to the job the applicant is seeking, and does the applicant know it is being measured?
- Why was an applicant not selected for the job (can big data and the machine learning algorithm provide explanations)?
- How can applicants improve their skills and abilities to become more qualified upon retesting?

Many test developers collect data related to a range of applicant experiences and reactions. In addition to asking test takers to rate how interesting or engaging they found the test to be, they may use the dropout rate (i.e., the number of incomplete tests) as a proxy for applicant engagement. Employers confront several problems with using these metrics. First, the metrics used rarely incorporate the full range of organizational considerations. For example, test developers may ask job applicants if the games were engaging, but they seldom ask applicants whether they felt that job relevant KSAOs were being measured. Second, good comparative data are rarely available. Vendors may share statistics on applicant reactions to their innovative assessments as a marketing tool, but rarely compare those reactions to reactions associated with other tools. Third, although there have been analyses comparing applicant reactions across applicants and organizations (Hausknecht et al., 2004), in any given local setting, it can be difficult to know the range of applicant reactions and how relatively positive or negative they may be. Fourth, applicant reactions may be affected by how well or poorly candidates think they performed at the time of testing, something that may be substantially more difficult for them to gauge for novel games or any selection tools having no obviously correct answers. Fifth, whether they were given a job offer or not is a huge driver of applicant reactions, and organizations should therefore consider whether they want to know about applicant reactions before offers were extended or after (or both). Sixth, simply asking questions of applicants about their testing experience after the testing event may alter their perceptions of their experience. For example, asking applicants a question about the fairness or invasiveness of a video interview, a gaming tool, or data scraped from the Internet may stimulate reflection on the experience that they would not have engaged in otherwise. Research evaluating applicant experiences and the extent to which test takers perceive the selection procedure to be not only engaging and unobtrusive but also a fair and accurate measure of job related KSAOs is needed.

Impact of new forms of selection procedures on well-qualified candidates. Researchers have found tenuous relationships between applicant reactions and applicant behavior (Ryan & Huth, 2008) that have been described as the Achilles heel of personnel selection (Sackett & Lievens, 2008, p. 439). Nevertheless, many organizations remain concerned about the effect of their selection programs and processes on the type and quality of applicants they can attract. In the age of technologically enhanced selection tools combined with machine learning algorithms, it remains unclear how candidates will react when they learn that their selection was influenced by some unknown weighted combination of their facial expressions, voice quality, mouse clicks, and a wide range of other types of data that vary in terms of perceived job relevance - in addition to their MBA credentials from a top school. What we do know is that applicant reactions to selection procedures may be more important to today's organizations than early applicant reactions researchers ever realized, because applicants can influence others more strongly through social media posts (e.g., Twitter, Facebook, LinkedIn).

Questions:

- *How do applicants evaluate organizations that minimize time invested in personal candidate interactions (e.g., through use of chatbots or avatars)?*
- *What are applicant reactions to various forms of innovative approaches to selection, and how do those reactions affect an employer's ability to attract qualified candidates and its own reputation?*
- *What is the effect of new forms of selection procedures on well-qualified candidates?*

Communications

Managers, whose success depends on a competent workforce, are particularly concerned that the skills critical to performing the job are being measured. In fact, perceived deficiencies in skill sets are often the impetus for new test development and validation projects. Additionally, labor organizations and advocacy groups for subsets of the applicant population have been particularly concerned about aspects of the selection procedures, including job relevance and fairness. Clearly, enforcement officials have a statutory and regulatory interest in what is being measured and how.

Sharing selection procedure information. Selection programs have always been of interest to applicants, organizations, employment agencies, career counselors, policymakers, and other stakeholders. A key question related to applicant reactions involves what to say about how people were selected to those who have an interest in the selection approach. There have always been some limits on what is shared and what information various stakeholders are entitled to, interested in, and can understand. Most organizations are unwilling to share anything that would jeopardize the continued use of the tests (e.g., specific item content or scoring keys) or increase the likelihood of a legal or administrative challenge (e.g., adverse impact data). The technical aspects of evaluating measurement or predictive bias are generally beyond the comprehension of most applicants and hiring managers (e.g., regression slopes and factor loadings); and sources of bias in machine learning models will be even more challenging to explain. Still, most applicants want to know at a basic level what KSAOs are being measured, how they are being evaluated, and, if they are not successful, when they can try again and what they can do to improve the next time.

Test preparation materials typically describe the selection process, state the logistics of administration, provide tips for preparing, and sometimes offer practice questions. In the AI domain, it is not clear what kinds of preparation materials could be given to applicants when the selection process will be based on face or voice characteristics; information extracted from resumes, applications, or social media; or performance in a group-based game scenario. There is a largely unresolved issue about whether training for a video interview is possible, and if so, whether it produces invalid variance (faking/lying) or valid variance by ensuring candidates understand what is expected of them in the interview. Some organizations sidestep the specific questions about a selection process by simply informing candidates that the selection outcome indicates whether or not they met the employer's requirements. Nonetheless, they may still receive questions from new hires and rejected applicants about why they were or were not hired.

Questions:

- *What information can and should be provided to unsuccessful applicants?*
- *What aspects of a selection procedure should an organization share with a range of other stakeholders (e.g., manager, industry, clients, customers, shareholders)?*

Ethics

As noted above, SIOP members are bound by APA's *Ethics Code* (2010). A review of this document, especially Section 9 that focuses on assessment, suggests several possible ethical concerns related to the use of some newer forms of assessment. Importantly, psychologists are required to base their opinions on information and techniques sufficient to justify their findings, use instruments with established evidence for reliability and validity (even as that evidence continues to be updated), and obtain informed consent that includes an explanation of the nature of the assessment:

> *9.01 Bases for Assessments*
> (a) Psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings.
> *9.02 Use of Assessments*
> (b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation.
> *9.03 Informed Consent in Assessments*
> (a) Psychologists obtain informed consent for assessments, evaluations, or diagnostic services, as described in Standard 3.10, Informed Consent, except when (1) testing is mandated by law or governmental regulations; (2) informed consent is implied because testing is conducted as a routine educational, institutional, or organizational activity (e.g., when participants volun-

> tarily agree to assessment when applying for a job); or (3) one purpose of the testing is to evaluate decisional capacity. Informed consent includes an explanation of the nature and purpose of the assessment, fees, involvement of third parties, and limits of confidentiality and sufficient opportunity for the client/patient to ask questions and receive answers. (APA, 2010)

Thus, reliability, validity, and fairness are not separate concerns but are intertwined and related to critical ethical concerns. When working in the testing and assessment arena, I-O psychologists must determine whether these ethical standards are being met or can be met. Psychologists must establish validity and reliability of the instruments they use for selection and have the evidence to support the recommendation made from the test score. If a machine learning algorithm is used to infer which job applicants will have a higher likelihood of good job performance, what is the quality and strength of the evidence to support this inference? In the end, it is the professional responsibility of I-O psychologists to require information about reliability, validity, and fairness as input for deciding whether a selection system—technology enhanced or otherwise—can be used to make inferences about job performance (or other criteria).

As noted in the discussion above, the ethics of using data over which the candidate has little, or no, control are hazy. Although informed consent is implied when testing is conducted as a routine part of the selection process, it is not clear if this implied consent applies only to a formal testing situation or also to data that the candidate might not be aware are being obtained and used. Note also that the ethical standards for informed consent are different for researchers (see Section 8.05 of the Code) developing selection tools than for those employing them. Guzzo et al. (2015) discuss the implications of these requirements in some detail (see also Dekas & McCune, 2015 for further implications).

A Call to Action

One of the purposes of this paper is to call the profession of I-O psychology to action in considering how the standards set in the *Principles for the Validation and Use of Personnel Selection Procedures* (*Principles*, SIOP, 2018) apply to technologically enhanced assessments used for employment decisions. To be clear, the *Principles* reflects the established science and practice of selection to date; it is not our recommendation to rewrite the *Principles*. Rather, our desire is to see interpretive guidance applying the *Principles* to technologically enhanced assessments that guides developers and users of employment tests in best practices and addresses the questions asked in this paper. Although there are different approaches to this task, we believe that our professional organization, SIOP, should sponsor this effort.

I-O psychologists are well-equipped to undertake this task, as many are trained extensively in the areas of measure development, psychometrics, personnel selection, and relevant employment law, and have deep experience in developing, validating, and managing the implementation of selection procedures in organizations. We have a deep grounding in factors that are critically important for employment testing, such as psychological constructs (e.g., knowledge, personality, interests, engagement, teamwork, safety, performance, turnover), theories of testing and assessment (e.g., construct-oriented test development, psychometric modeling, appropriate scoring and interpretation), the types of evidence that support the inferences to be made from the test scores (e.g., selection decisions, validity), psychometric properties of effective tests (e.g., internal consistency, test–retest, and alternate forms reliability), and the evaluation of subgroup differences (e.g., differential prediction, measurement invariance, and adverse impact with respect to protected classes). In addition, SIOP has a long history of documenting the consensus of opinion on research and practice in employment testing in the *Principles*.

However, our knowledge and experience must be supplemented with that of others who work in this field. Data scientists and software developers are important collaborators in developing technologies to acquire, store, and analyze large amounts of information, create algorithms that predict outcomes, and evaluate their effectiveness. Web designers and IT professionals are needed to construct games and create engaging and effective web interfaces, producing tools to be used by applicants and interpreted by recruiters, HR professionals, I-O psychologists, and hiring managers. In addition, the legal profession in the U.S. has a deep interest in how new selection procedures comply with existing federal, state, and local laws as well as meet the requirements of regulatory agencies.

As I-O psychologists, we cannot regulate the practice of others; however, many of us serve as experts advising organizations and the government and testifying with respect assessments (both supporting and challenging). Guidance in interpreting the *Principles* would clarify expectations and provide needed consistency. Our suggestion for collaborative development of interpretive guidelines across and an array of other disciplines is intended to be help fill knowledge gaps among participating parties.

We believe that we must engage in conversation and debate with professionals in applied statistics, computer science, and other disciplines to learn about their applications of new machine learning methodologies. Together, we can better identify the strengths, critique the weaknesses, and understand appropriate and inappropriate applications.

Conclusion

New technologically enhanced assessments present opportunities to broaden selection procedures and make them more efficient. At the same time, current practices in this area come with some serious liabilities and potential risks, that must be addressed through the lens of professional guidelines, expertise, and experience of I-O psychologists who work in the field of employee selection. It is our hope that more I-O psychologists will proactively engage in this assessment arena (not only selection specialists, but also in collaboration with those involved in recruiting, diversity and inclusion, and leadership), because it offers the possibility of improving assessment and promoting the future relevance of our profession. We believe that I-O psychologists also must vigorously engage with others who work in this area.

The work being done by I-O psychologists and others in the development of new assessment and selection tools is exciting and offers advantages to employers and applicants alike. Yet, we are also responsible for ensuring that progress does not approach escape velocity from our moorings of scientific, psychometric, and practical knowledge; understanding of legal guidelines, professional and ethical obligations; and many hard lessons learned in the employment testing arena. Now is the time to carefully consider how the *Principles* should be applied to new and evolving forms of assessments to reflect the research literature and best practices.

REFERENCES

Aiken, J.R., & Hanges, P.J. (2015). Teach an I-O to fish: Integrating data science into I-O graduate education. Industrial and Organizational Psychology, 8(4), 539-544.

Ajunwa, I. (2019, October 8). Beware of automated hiring. The New York Times. Retrieved from https://www.nytimes.com/2019/10/08/opinion/ai-hiring-discrimination.html.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Psychological Association. (2010). Ethical principles of psychologists and code of conduct. Washington, DC: American Psychological Association.

Arthur, Jr., W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. Journal of Applied Psychology, 93(2), 435-442. https://doi.org/10.1037/0021-9010.93.2.435

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. Journal of Applied Psychology, 100(2), 431–449. https://doi.org/10.1037/a0038047

Braun, M.T., & Kuljanin, G. (2015). Big data and the challenge of construct validity. Industrial and Organizational Psychology, 8(4), 521-527.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105. https://doi.org/10.1037/h0046016

Civil Rights Act of 1964 § 7, 42 U.S.C. §2000e et seq (1964). Retrieved from Equal Employment Opportunity Commission website: http://www.eeoc.gov/laws/statutes/titlevii.cfm.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Converse, P. D., & Oswald, F. L. (2014). Thinking ahead: Assuming linear versus nonlinear personality-criterion relationships in personnel selection. Human Performance, 27(1), 61–67.

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. Journal of Applied Psychology, 75(3), 297–300.

Cowen, A. S., & Keltner, D. (2019). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. American Psychologist, 75, 349-364. https://doi.org/10.1037/amp0000488

Dekas, K., & McCune, E. A. (2015). Conducting ethical research with big and small data: Key questions for practitioners. Industrial and Organizational Psychology, 8(4), 563-567.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87. https://doi.org/10.1145/2347736.2347755

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38290 38315.

Equal Employment Opportunity Commission. (1979). Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. Federal Register, 44 (43) . https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines xx-xx.

Guardians Association of New York City Police Department, Inc. v. Civil Service Commission of City of New York, 630 F.2d 79 (2d Cir. 1980))

Guion, R. M. (2011). Assessment, measurement, and prediction for personnel decisions. (2nd ed.) New York, NY: Routledge. https://doi.org/10.4324/9780203836767

Gutman, A., Koppes, L.L., & Vodanovich, S.J. (2011). EEO law and personnel practices (3rd ed.). New York: Routledge.

Guzzo, R.A., Fink, A.A., King, E., Tonidandel, S., & Landis, R.S. (2015). Big data recommendations for industrial-organizational psychology. Industrial and Organizational Psychology, 8(4), 491-508.

Harwell, D. (2019, October 22). A face-scanning algorithm increasingly decide whether you deserve the job. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Harwell, D. (2019, November 6). Rights group files federal complaint against AI-hiring firm HireVue, citing "'unfair and deceptive' deceptive" practices. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. Personnel Psychology, 57(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. Personnel Psychology, 49, 51-83. https://doi.org/10.1111/j.1744-6570.1996.tb01791.x

Illingworth, A.J. (2015). Big data in I-O psychology: Privacy considerations and discriminatory algorithms. Industrial and Organizational Psychology, 8(4), 567-575.

Kaleghi, B. (2019, July). The how of explainable AI: Post-modeling explainability. Towards Data Science. Retrieved from https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98(6), 1060–1072.

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. Journal of Applied Psychology, 102(1), 43–66. https://doi.org/10.1037/apl0000160

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology, 60(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

Morgeson, F. P., Brannick, M. T., & Levine, E. L. (2020). Job and work analysis: Methods, research, and applications for human resource management (3rd ed.). Sage.

Narayanan, A. (2019, Apr. 22). How to recognize AI snake oil [PowerPoint presentation]. Arthur Miller Lecture on Science and Ethics, Cambridge, MA, United States. https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

OFCCP (2019). Validation of employee selection procedures. https://www.dol.gov/agencies/ofccp/faqs/employee-selection-procedures#Q6. Retrieved April 10, 2020.

Oswald, F. L. (2020). Future research directions for big data in psychology. In S. E. Woo, L. Tay, & R. Proctor (Eds.). Big data in psychological research (pp. 427-441). Washington, DC: APA Books.

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resources management: Forward progress for organizational research and practice. Annual Review of Organizational Psychology and Organizational Behavior, 7, 505-533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

Oswald, F. L., & Putka, D. J. (2016). Statistical methods for big data: A scenic tour. In S. Tonidandel, E. B. King, & J. M. Cortina (Eds.), SIOP organizational frontier series. Big data at work: The data science revolution and organizational psychology (p. 43–63). Routledge/Taylor & Francis Group.

Oswald, F. L., & Putka, D. J. (2017). Big data methods in the social sciences. Current Opinion in Behavioral Sciences, 18, 103-106. https://doi.org/10.1016/j.cobeha.2017.10.006

Polli, F. (2019, October 29) Using AI to eliminate bias from hiring. Harvard Business Review. Retrieved from https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring.

Pyburn, K. M., Jr., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. Personnel Psychology, 61(1), 143–151. https://doi.org/10.1111/j.1744-6570.2008.00108.x

Raghavan, M., Barocas, S., Keinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Roth, P. L., Thatcher, J. B., Bobko, P., Matthews, K. D., Ellingson, J. E., & Goldberg, C. B. (2020). Political affiliation and employment screening decisions: The role of similarity and identification processes. Journal of Applied Psychology, 105(5), 472–486. https://doi.org/10.1037/apl0000422

Rotolo, C.T., & Church, A. H. (2015). Big data recommendations for industrial-organizational psychology: Are we in Whoville? Industrial and Organizational Psychology, 8(4), 515-520.

Ryan, A.M., & Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. Human Resource Management Review, 18 (3), 119-132. ISSN 1053-4822, https://doi.org/10.1016/j.hrmr.2008.07.004.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. Annual Review of Psychology, 59, 419–450. https://doi.org/10.1146/annurev.psych.59.103006.093716

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262–274.

Schroeder, A. N., & Cavanaugh, J. (2018). Fake it 'til you make it: Examining faking ability on social media pages. Computers in Human Behavior, 84, 29-35. https://doi.org/10.1016/j.chb.2018.02.011

Singer, N., & Metz, C. (2019, December 19). Many facial-recognition systems are biased, says U.S. study. The New York Times. Retrieved from https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html.

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures (5th ed.). Retrieved from https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf

Tippins, N. T. (2009). Where is the unproctored Internet testing train headed now? Industrial and Organizational Psychology: Perspectives on Science and Practice, 2(1), 69–76. https://doi.org/10.1111/j.1754-9434.2008.01111.x

Tippins, N. T., & Adler, S. (2011). Technology-enhanced assessment of talent. New York, NY: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118256022

Totty, M. (2020, November 3). How to make artificial intelligence less biased. Wall Street Journal. Retrieved from https://www.wsj.com/articles/how-to-make-artificial-intelligence-less -baised-11604415654.

Walmsley, P. T., Sackett, P. R., & Nichols, S. B. (2018). A large sample investigation of the presence of nonlinear personality-job performance relationships. International Journal of Selection and Assessment, 26(2-4), 145-163. https://doi.org/10.1111/ijsa.12223

Weed, J. (2020, November 27). Job interviews without the interviewers, products of the pandemic. The New York Times. Retrieved from https://www.nytimes.com/2020/11/27/business/video-job-intrviews.html?referringSource=articleShare.

Zhang, L., Van Iddekinge, C. H., Arnold, J. D., Roth, P. L., Lievens, F., Lanivich, S. E., & Jordan, S. L. (2020). What's on job seekers' social media sites? A content analysis and effects of structure on recruiter judgments and predictive validity. Journal of Applied Psychology. Advance online publication. https://doi.org/10.1037/apl0000490

RECEIVED 02/16/21 ACCEPTED 05/21/21